REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
28-08-2010	Final Report	1-Oct-2009 - 30-Jun-2010

4. TITLE AND SUBTITLE
Information Assurance: Detection & Response to Web Spam Attacks

5a. CONTRACT NUMBER
W911NF-09-1-0566

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Pang-Ning Tan, Anil K Jain

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Michigan State University
Contract & Grant Admin.
Michigan State University
East Lansing, MI 48824 -

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
56802-CS-II.1

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
As online social media applications such as blogs, social bookmarking (folksonomies), and wikis continue to gain its popularity, concerns about the rapid proliferation of Web spam has grown in recent years. These applications enable spammers to submit links that divert unsuspected users to spam Web sites. The goal of

15. SUBJECT TERMS
information assurance, web spam, social media

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Pang-Ning Tan
UU	UU	UU	UU		19b. TELEPHONE NUMBER: 517-432-9240

Report Title

Information Assurance: Detection & Response to Web Spam Attacks

ABSTRACT

As online social media applications such as blogs, social bookmarking (folksonomies), and wikis continue to gain its popularity, concerns about the rapid proliferation of Web spam has grown in recent years. These applications enable spammers to submit links that divert unsuspected users to spam Web sites. The goal of this research is to investigate novel techniques to detect Web spam in social media web sites. Specifically, we have developed a co-classification framework that simultaneously detects web spam and the spammers who are responsible for posting them on social media web sites. Using data from two real-world applications, we empirically showed that the proposed co-classification framework is more effective that learning to classify the Web spam and spammers independently. We also investigated an approach to enhance the framework by leveraging out-of-domain data collected from multiple social media web sites.

List of papers submitted or published that acknowledge ARO support during this reporting period. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Number of Papers published in peer-reviewed journals: 0.00

(b) Papers published in non-peer-reviewed journals or in conference proceedings (N/A for none)

Number of Papers published in non peer-reviewed journals: 0.00

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

P.-N. Tan, F. Chen, and A.K. Jain. Web spam: A case of misinformation in online social networks. In \emph{Proceedings of the Workshop on Information in Networks (WIN-2009)}, New York, 2009.

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts): 1

Peer-Reviewed Conference Proceeding publications (other than abstracts):

F. Chen, P.-N. Tan, and A.K. Jain. A co-classification framework for detecting Web spam and spammers in social media Web sites. In Proceedings of the Conference on Information and Knowledge Management (CIKM-2009), Hong Kong, 2009.

P.-N. Tan, F. Chen, and A.K. Jain. Information assurance: Detection of Web spam attacks in social media. In Proceedings of the 27th Army Science Conference, Orlando, FL, 2010.

L. Liu and P.-N. Tan. A Framework for Co-Classification of Articles and Users in Wikipedia. To appear in Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence (WI-2010), Toronto, Canada, 2010.

P. Mandayam Comare, P.-N. Tan, and A. K Jain. Multi-task Learning on Multiple Related Networks. To appear in Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM 2010), Toronto, Canada (2010).

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts): 4

(d) Manuscripts

Number of Manuscripts: 0.00

Patents Submitted

Patents Awarded

Graduate Students

NAME	PERCENT_SUPPORTED
Prakash Mandayam Comare	0.50
Feilong Chen	0.00
FTE Equivalent:	**0.50**
Total Number:	**2**

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Pang-Ning Tan	0.00	No
Anil K Jain	0.00	No
FTE Equivalent:	**0.00**	
Total Number:	**2**	

Names of Under Graduate students supported

NAME	PERCENT SUPPORTED

FTE Equivalent:

Total Number:

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period:	0.00
The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:.....	0.00
The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:.....	0.00
Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):	0.00
Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:	0.00
The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense	0.00
The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:	0.00

Names of Personnel receiving masters degrees

NAME

Total Number:

Names of personnel receiving PHDs

NAME

Total Number:

Names of other research staff

NAME	PERCENT SUPPORTED

FTE Equivalent:

Total Number:

Sub Contractors (DD882)

Inventions (DD882)

FINAL REPORT

on

ARO Grant Number W911NF-09-1-0566

(October 1, 2009-June 30, 2010)

Information Assurance: Detection & Response to Web Spam Attacks

Project URL: http://www.cse.msu.edu/∼ptan/project/webspam/webspam.html

Investigators:

Dr Pang-Ning Tan
Dept. of Comp Science and Engineering
Michigan State University
3115 Engineering Building
East Lansing, MI 48824
Tel: 517-432-9240
Fax: 517-432-1061
Email: ptan@cse.msu.edu

Dr Anil K Jain
Dept. of Comp Science and Engineering
Michigan State University
3115 Engineering Building
East Lansing, MI 48824
Tel: 517-432-9240
Fax: 517-432-1061
Email: jain@cse.msu.edu

Foreword

As online social media applications such as blogs, social bookmarking (folksonomies), and wikis continue to gain its popularity, concerns about the rapid proliferation of Web spam has grown in recent years. These applications enable spammers to submit links that divert unsuspected users to spam Web sites. The goal of this research is to investigate novel techniques to detect Web spam in social media web sites. Specifically, we have developed a co-classification framework that simultaneously detects web spam and the spammers who are responsible for posting them on social media web sites. Using data from two real-world applications, we empirically showed that the proposed co-classification framework is more effective that learning to classify the Web spam and spammers independently. We also investigated an approach to enhance the framework by leveraging out-of-domain data collected from multiple social media web sites.

Contents

1 Statement of the Problem Studied

The explosive growth of the Internet has transformed the way we communicate and interact with each other. The Internet, which was once the realm of email, FTP, and Usenet, is barely recognizable nearly two decades later with the emergence of social media applications such as weblogs, wikis, twitters, folksonomies, and video or photo file sharing sites. Instead of passively searching and consuming information, users nowadays are actively engaged in the creation and distribution of information using tools provided by the social media Web sites. These tools often allow users to submit links to interesting online articles or add shortcuts (bookmarks) to their favorite Web sites. The emergence of social media applications has led to growing concerns about the alarming increase of Web spam as spammers may exploit the capabilities provided by these applications to submit links that direct users to spam Web sites. Worse still, some of the directed Web sites may trick unsuspected users into divulging their personal information or allow malicious code to be injected to the user's browser. To alleviate such Web spam attacks, it is therefore critical to develop effective techniques that can automatically detect Web spam and spammers in social media applications.

This report begins with our investigation into the prevalence and characteristics of Web spam at two popular social media Web sites, delicious.com and digg.com [10]. We then present a novel learning paradigm called co-classification to simultaneously detect Web spam and spammers based on their content and link information [3]. We also investigate the effectiveness of augmenting data from multiple social media applications to improve Web spam detection using a combination of co-training with the co-classification approach [11]. We also investigate extensions of the co-classification framework to other network classification problems [7, 4].

2 Summary of the Most Important Results

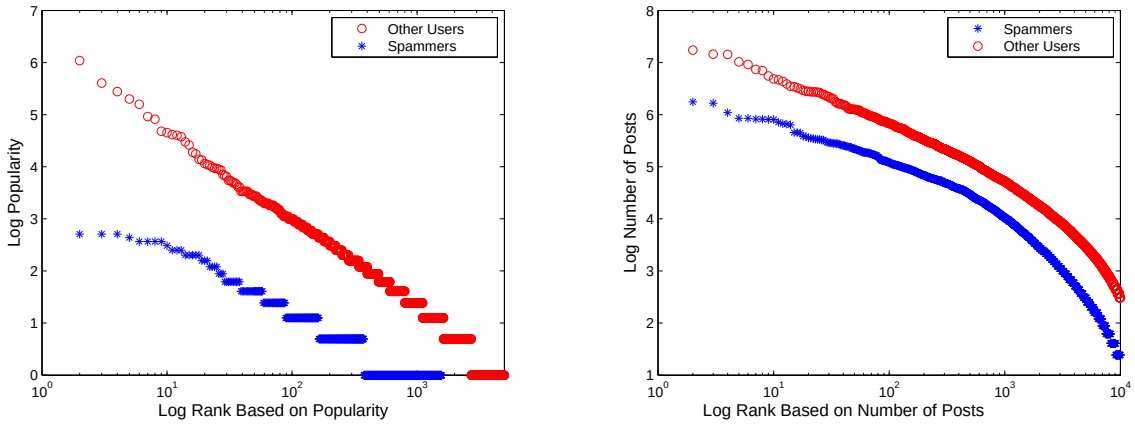
2.1 Web Spam in Social Media Web Sites

In [10], we analyzed the prevalence and characteristics of Web spam at two popular social media Web sites, delicious.com and digg.com. The former is a social bookmarking Web site that allows users to add shortcuts (bookmarks) to the URLs of their favorite Web sites, assign tags to each bookmark, and share them with other users. The latter is a social news Web site, which allows users to post links to interesting news stories they found on the Internet or vote on the stories submitted by other users. Using a list of spam Web sites extracted from a benchmark corpus [12], nearly 7% of them were found posted at digg.com and 18% of them at delicious.com. These results showed the prevalence of Web spam in social media and suggested the need for automated tools to detect them in order to improve quality of online information and to prevent unsuspected users from being diverted to spam and other malicious Web sites.

Although some social media applications such as `digg` provide additional counter-measures to safeguard against the promotion of Web spam (e.g., by allowing users to "vote down" or "bury" uninteresting posts), these measures are not entirely full proof because spammers may create several bogus user accounts and collude with each other to promote ("vote up" or "dig") their spam Web sites. The problem is even more acute at delicious.com, where nearly one-third of the spam URLs have been bookmarked by at least 20 users and about 23% of them were bookmarked by at least 30 users. Some of the spam URLs were as popular as the non-spam URLs listed at http://delicious.com/popular/. An example of a popular spam URL at delicious.com was the `Airset spam`, which was initially discovered by Brian Dear[1]. He noted several unusual characteristics of the Airset spam, including: (1) all the bookmarks correspond to the same URL, (2) all the bookmarks were assigned the same keyword tag EVDB, and (3) the majority of users who submitted the spam URL posted no other URLs. While such an unusual pattern is a potentially useful signature for Web spam, it is insufficient to uncover all types of spam as the more experienced spammers may submit links to other legitimate Web sites to obfuscate their spamming activities.

To illustrate the difficulty in identifying Web spam and spammers, consider the plots shown in Figure 1. Figure 1(a) compares the user popularity for spammers against non-spammers at delicious.com. User popularity refers to the number of "fans" who subscribe to a user's network. Although their scales are quite different, i.e., the most popular spammers have fewer fans than the most popular non-spammers, both plots appear to exhibit a power law distribution. In terms of the number of URLs submitted by spammers and non-spammers, again, the shape and amplitude of the distributions are close to each other, as shown in Figure 1(b). This observation suggests that user popularity and their number of posted bookmarks are not sufficient to effectively detect Web spam and spammers. This is because it would be difficult to set an appropriate minimum popularity or number of posted bookmarks threshold to filter the spammers and spam URLs without misclassifying the non-

[1] A discussion of the Airset spam can be found at http://www.brianstorms.com/archives/000575.html.



(a) Distribution of user popularity

(b) Distribution of number of posts submitted

Figure 1: Comparing the user popularity and number of posts submitted by spammers against non-spammers at delicious.com social media Web site.

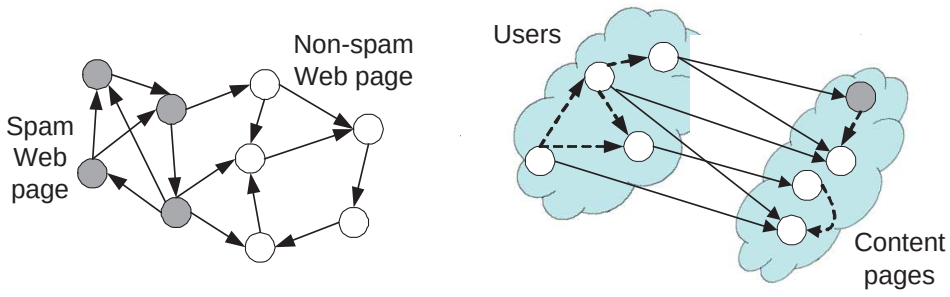
spammers and non-spam URLs. We need to consider other link-based and content-based features to improve the detection rate of Web spam and spammers.

2.2 Co-Classification Framework for Web Spam Detection

While there has been extensive research on detecting spam on the World Wide Web [8, 9, 5, 6, 2, 1], spam detection in social media is still in its infancy. Figure 2 illustrates the conceptual difference between spam detection on the World Wide Web and spam detection in social media applications. The former is composed of a single, homogeneous network consisting of nodes of the same type (Web pages) while the latter is a multi-graph network containing nodes of different types (users and their submitted URLs). Given the nature of the data, spam detection for social media applications can be decomposed into two sub-problems, namely, detecting spam URLs and the spammers who are responsible for posting them.

There are many types of features that can be used for Web spam detection in social media. For example, content-based features can be derived from the text description and tags assigned by users to the URLs they have submitted. Link-based features can also be constructed from the links between users, links between URLs, or links between users and their submitted URLs. However, integrating such diverse features into a Web spam detection algorithm is not a trivial task. First, existing classifiers such as support vector machine (SVM) are not designed to handle both content-based and link-based features. Second, the links are often noisy due to the fact that some legitimate users may inadvertently link to spam URLs whereas some spammers may deliberately post links to legitimate Web sites to evade detection.

In [3], we have developed a robust framework to effectively detect Web spam and spammers in social media Web sites. Our framework extends the least-square support vector



(a) Spam detection in World Wide Web (b) Spam detection in social media

Figure 2: Comparison between spam detection in the World Wide Web (where the network consists of hyperlinked Web pages) and spam detection in social media (where the network consists of users and their shared social media content).

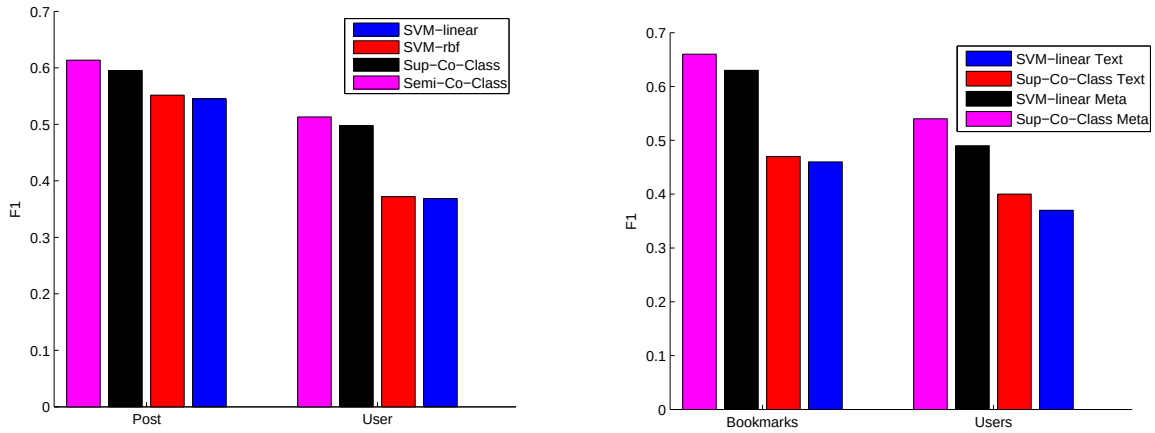
machine (LS-SVM) classifier to handle data that contains both link-based and content-based features. The framework was developed based on the following two assumptions: (1) Spam URLs are more likely to be posted by spammers than non-spammers and (2) Spammers are more likely to link to other spammers than to non-spammers. We formalize these assumptions as graph regularization constraints and develop a co-classification algorithm to learn a pair of classifiers that simultaneously detect Web spam and spammers at a social media Web site. We also showed that our co-classification framework can be extended to nonlinear models using the kernel trick and adapted to a semi-supervised learning setting.

Figure 3 shows the results of detecting Web spam and spammers at delicious.com and digg.com Web sites. The results indicate that our supervised and semi-supervised co-classification algorithms significantly outperform techniques that learn to classify the Web spam and spammers independently. In addition, the semi-supervised co-classification algorithm was more effective than the supervised version. This is because the semi-supervised algorithm takes advantage of the link information to propagate the labeled information to neighboring nodes (users and URLs).

2.3 Web Spam Detection with Out-of-Domain Data

One of the challenges in Web spam detection for social media applications is that training examples are often scarce and expensive to acquire. The proliferation of social media Web sites gives an opportunity to leverage data from different sources to improve model performance. For example, one may enhance the performance of a classifier constructed from `delicious.com` using out-of-domain data from `digg.com`. This is a reasonable assumption since the spam Web sites are often posted on different social media Web sites.

In [11], we have developed a method based on co-training to utilize out-of-domain data for improving Web spam detection. Co-training (Blum et. al., 1998) is a semi-supervised learning technique that assumes each data point can be represented by two disjoint sets of



(a) Performance comparison for delicious.com data (b) Performance comparison for delicious.com data

Figure 3: Comparison between the supervised and semi-supervised co-classification algorithms against SVM classifiers trained on the user and URL networks independently.

features. Each feature set provides a complementary view of the data point. Ideally, the two feature sets should be conditionally independent given the class. Furthermore, each feature set should contain relevant information to correctly predict the class label of a data point. If both conditions are satisfied, it can be shown that co-training will improve classification accuracy on the target domain.

Our proposed co-training with co-classification approach first learns an initial pair of classifiers for each domain source (digg.com and delicious.com). It then applies the classifiers to the test examples and selects the test examples with highest confidence in their predictions to be augmented to the labeled training data. This process is repeated until the algorithm converges. We evaluated the performance of our hybrid co-training with co-classification algorithm using the delicious.com and digg.com datasets. After checking the submitted URLs, we found about 8% of the URLs are common to both Web sites. In order to analyze the effect of using out-of-domain data, we gradually increase the proportion of common URLs in the training set from 4% to 8%. The experimental results given in Figure 4 showed that the performance of co-training with co-classification, denoted as `Co-Co-Class`, is better than applying co-classification on data from a single domain, especially when the proportion of common URLs posted on both Web sites increased.

2.4 Generalization of Co-Classification Framework

The original co-classification framework developed in [3] was designed for discriminating binary classes only. Since Web spam can be divided into different subclasses, it would be useful to extend the framework to more than two classes. In [7], we have generalized the co-classification framework to multi-class problems. Specifically, we formalized the joint classification tasks as a constrained optimization problem, in which the relationships between

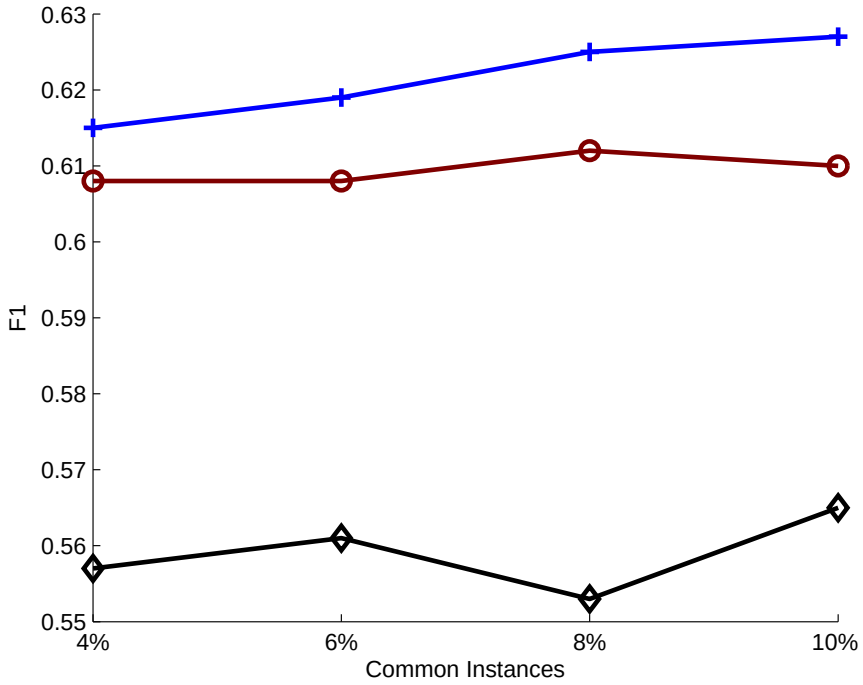


Figure 4: Performance comparison between co-classification on networks from a single domain against co-classification with co-training on network data from multiple domains.

the classes in two different networks are modeled as graph regularization constraints. Unlike our previous binary class formulation, our new approach also allows us to incorporate prior knowledge about the potential relationships between classes in different networks to avoid overfitting. Experimental results showed that the proposed algorithm significantly outperforms classifiers that learn each classification task independently.

The co-classification framework assumes that labeled examples are available on both user and URL networks. Thus, it is not applicable when labeled examples are available in only one of the two networks. In [4], we presented an approach for multi-task learning in multiple related networks, where in we perform supervised classification on one network and unsupervised clustering on the other. We showed that the framework can be extended to incorporate prior information about the correspondences between the clusters and classes in different networks. Through various set of experiments, we have demonstrated the effectiveness of the proposed framework compared to independent classification or clustering on individual networks.

3 List of Publications

Conference/Workshop Proceedings:

1. P.-N. Tan, F. Chen, and A.K. Jain. Web spam: A case of misinformation in online social networks. In *Proceedings of the Workshop on Information in Networks (WIN-2009)*, New York, 2009.

2. F. Chen, P.-N. Tan, and A.K. Jain. A co-classification framework for detecting Web spam and spammers in social media Web sites. In *Proceedings of the Conference on Information and Knowledge Management (CIKM-2009)*, Hong Kong, 2009.

3. P.-N. Tan, F. Chen, and A.K. Jain. Information assurance: Detection of Web spam attacks in social media. In *Proceedings of the 27th Army Science Conference*, Orlando, FL, 2010.

4. L. Liu and P.-N. Tan. A Framework for Co-Classification of Articles and Users in Wikipedia. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence (WI-2010)*, Toronto, Canada, 2010.

5. P. Mandayam Comare, P.-N. Tan, and A. K Jain. Multi-task Learning on Multiple Related Networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM 2010)*, Toronto, Canada (2010).

4 List of Project Participants

Co-Principal Investigators:

- Pang-Ning Tan
- Anil K Jain

Graduate Students:

- Feilong Chen
- Prakash Mandayam Comare

References

[1] L. Becchetti, C. Castillo, D. Donato, R. Baeza-YATES, and S. Leonardi. Link analysis for web spam detection. *ACM Trans. Web*, 2(1):1–42, 2008.

[2] A. Benczur, K. Csalogany, T. Sarlos, and M. Uher. Spamrank - fully automatic link spam detection. In *Proc. International Workshop on Adversarial Information Retrieval on the Web*, 2005.

[3] F. Chen, P.-N. Tan, and A. Jain. A co-classification framework for detecting web spam and spammers in social media web sites. In *Proc. of the Conference on Information and Knowledge Management (CIKM-2009)*, Hong Kong, 2009.

[4] P. M. Comare, P.-N. Tan, and A. Jain. Multi-task learning on multiple related networks. In *Proc. of the Conference on Information and Knowledge Management (CIKM-2010)*, Toronto, Canada, 2010.

[5] Z. Gyongyi, P. Berkhin, H. Garcia-Molina, and J. Pedersen. Link spam detection based on mass estimation. In *Proc. the 32nd international Conference on Very Large Data Bases*, 2006.

[6] Z. Gyongyi and H. Garcia-Molina. Web spam taxonomy. In *Proc. International Workshop on Adversarial Information Retrieval on the Web*, 2005.

[7] L. Liu and P.-N. Tan. A framework for co-classification of articles and users in wikipedia. In *Proc. of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence (WI-2010)*, Toronto, Canada, 2010.

[8] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *Proc. of the 15th Int'l Conf on World Wide Web*, pages 83–82, 2006.

[9] R. Raj and V. Krishnan. Web spam detection with anti-trust rank. In *Proc. 2nd International Workshop on Adversarial Information Retrieval on the Web*, 2006.

[10] P.-N. Tan, F. Chen, and A. Jain. Web spam: A case of misinformation in online social networks. In *Proc. of the Workshop on Information in Networks (WIN-2009)*, New York, 2009.

[11] P.-N. Tan, F. Chen, and A. Jain. Information assurance: Detection of web spam attacks in social media. In *Proc. of the 27th Army Science Conference*, Orlando, FL, 2010.

[12] S. Webb, J. Caverlee, and C. Pu. Introducing the webb spam corpus: Using email spam to identify web spam automatically. In *Proc. of CEAS '06*, 2006.